Swirlifying Stats Training: Facilitating the Transition to R

Christopher K. Butler
The University of New Mexico
Albuquerque, NM

February 5, 2020

Abstract

What works well in teaching statistics and statistical programming to a novice group of students? I present key insights from more than a decade of teaching statistical methods with constant tinkering. These boil down to having frequent online quizzes with unlimited attempts and providing starter scripts with embedded tasks for each lab. Because some changes took effect during the same semester and cohorts vary considerably, this is not a cut-and-dried template for success. However, assessment data is presented to support the value of online quizzes with unlimited attempts. The paper also includes a discussion of transitioning the class from Stata to R and creating interactive tutorials (using the Swirl library in R) to aid learning.

NOTE: My apologies for the Frankensteinian nature of this paper. It is cobbled together from my winning Presidential Teaching Fellow application, an internally funded teaching-allocation grant application and follow up report, a presentation prepared for UNM's Center for Teaching and Learning on using online quizzes, and new material for the APSA 2020 Teaching and Learning Conference.

1 Introduction

We learn from doing, but we often learn faster from failing. Creating *an environment for safe interaction and repeated feedback* produces genuine learning and, in turn, successful students. Part of this environment is easier to create due to technology. Online quizzes and tutorials that give automated feedback remove some social and psychological stigma of being told we're something "wrong". But it is equally important to create a classroom environment where students can see the bar that has been set, they can see the ladder that will help them get over the bar, and they faith that the instructor is there to help them up each rung of the ladder.

There is a phrase we use in assessment workshops that struck a chord with me: "I can teach a dog to whistle." We use this phrase to distinguish between *teaching* and *learning*. Sure, I can show a dog how I whistle; I can explain to a dog the principles of whistling; but no matter how much teaching I do, the dog cannot whistle. The first time I saw this phrase in an assessment workshop, I was actually appalled. "Students are not like dogs!", I argued. "Students are *capable* of learning what we teach!" My emotional reaction reveals my bedrock teaching philosophy: everyone can learn what others have already learned (not just what I teach). But now I take a second, subtler lesson from this phrase. We need to gauge what the students are capable of presently and lead them to a higher level of achievement. This is the learning process—for both the students and the teacher.

This distinction between *teaching* and *learning* is equally important when creating an environment for safe interaction and repeated feedback. The active-learning critique of traditional lecture teaching is that students are not sufficiently engaged. Some will indeed learn through (or after) a traditional lecture as they engage with the material on their own. The goal of active learning is to increase the breadth and depth of engagement. But true engagement entails trying and failing and trying again (and sometimes failing again) until some level of success is achieved. Even when succeeding on the first try, it is important to try again to verify genuine understanding. Thus, I focus on repeated *feedback* rather than failure.

My current teaching methods emphasize using quizzes as a learning tool (rather than merely for evaluation), providing clear expectations about assignments from the beginning of the semester, and breaking down long-term projects into phases or other components.[1] Of equal importance here is an

[1] These methods are in addition to active-learning strategies that I adopted long ago, including in-class activities, small group

attempt to link the methods together. For example, quiz material may represent one theoretical application that the students would need to use in their final papers. Allowing them to retake quizzes with the aim of learning and mastery provides a foundation from which they can more readily succeed on that learning objective in another setting, such as their final papers.

Before discussing specific details of my instruction, I'd like to take a moment to discuss assessment of student learning and how it has changed my teaching methods. Evaluation of my students' learning is as much a part of my philosophy and method as is anything else. Even before I was introduced to the "learner-centered" instruction I've come to embrace (and the associated assessment of student learning objectives), I would disaggregate my students' performance is various ways, whether by content areas (such as security, political economy, and international law), question types (such as multiple choice, short answer, and maps), or sections of a final paper (such as literature, theory, and data analysis). Such disaggregation of students' performance told me where they were weakest and, hence, where I should consider how to change my instruction so the next class would perform better than the current class.[2] Little did I know that I had been doing a form of assessment of student outcomes since I had first been teaching.

The important shift, of course, was measuring specific learning objectives rather than the hodge podge disaggregation I had engaged in previously. The hardest part of this shift in instruction for me was creating assignments and exam questions that aligned with learning objectives. From there, it finally became self-evident to construct the course schedule as a *ladder of learning* upon which the students would "level up" as they demonstrated their achievements.[3] This has transformed the way I structure classes, grading, and assignments. I seek to show the students what high-level achievement looks like from the beginning of the semester by telling them how they will be evaluated on their final project. This represents my high bar for the students, but I'm showing them how high it is at the outset. Next, I show them how they can attain the skills they need to reach that high bar. This represents the rungs on the ladder of learning on which the

discussions or working groups, and, when appropriate, feeding the ducks at the Duck Pond.

[2] I have used assessment data on a quicker improvement schedule. For example, I taught a four-week online topics class called Political Games in which students work through four modules each week, each module having a quiz. Before the next week's modules open, I would evaluate what the students did not fully understand by conducting an item analysis of the quiz responses. I would then record a new voice-over-whiteboard lesson that reviewed these areas of weakness, presenting them in a different and/or clearer way. The students would then need to pass a quiz specifically designed to assess these areas before moving on to the new modules.

[3] I also have to thank my elder son for sending me a thought-provoking link on "Gamifying Education". (See Extra Credits: Gamifying Education on YouTube.).

students will pull themselves up to that high bar. Some students might be able to skip a few rungs; others will struggle to climb up each rung. This is my way of figuring out where the students are—collectively and individually—and how they're progressing along the pathway for achievement.

Now, for all my talk of learner-centered instruction, you might expect me to paint a perfectly rosy picture of continuous improvement. But assessment data are rarely so clear, let alone so rosy. What I have found instead are a few types data patterns where I can connect the data with instructional changes. I enumerate only a few of these patterns here. First, there are "trade-off" changes in which students perform better in areas I have focused instructional effort but perform worse in areas where I spent less time as a consequence. These changes need to be evaluated as to whether there was a net improvement across the trade-off. Second, there are the "better-luck-next-time" changes where the data suggest poorer performance despite instructional change but you really think the instructional change—perhaps with a clearer presentation or better execution—is worth a second chance. And third, there are "the hard nuts to crack" or areas where students continue to evaluate poorly despite instructional changes aimed to improve learning in those areas. Theses are areas where we just keep scratching our heads and trying new things. I'll discuss these three patterns in my detailed discussion below.

2 Course Aims and Objectives

Statistics for Social Research is our department's introductory graduate statistics course. As such, it is an excellent example of needing to figure out where the students stand and bringing them all up to a high level of achievement. In a way, I have set numerous high bars for this class: They will learn the basics of statistical inference, but they will also learn about computing for research, and how to write for a technical, research audience. This is a class I have taught for a decade. Even so, I continue to seek how to make it a better course.

In the past five years, the R programming language has become the tool of choice for quantitative scholars from a wide variety of disciplines. Unlike statistical analysis software like Stata or SPSS, R is free, open source, and compatible with all types of computing platforms (such as Windows, MacOS, or Linux). Programming in R is a common sought skill of data science and statistical analysis jobs in the private sector. In fact, it is the fifth-most mentioned skill for data science jobs, ranked well ahead of other statistical platforms and behind only high-level computer science programming languages like Python or Java (see Figure 1). It has also enjoyed significant growth in cited works from peer-reviewed academic journals (see Figure 2). As part of this trend, our most recent faculty hire and the most competitive job candidates are more fluent in R than in Stata (if they were taught Stata at all). As the next generation of faculty work with our graduate students, faculty will expect their training to be in R.

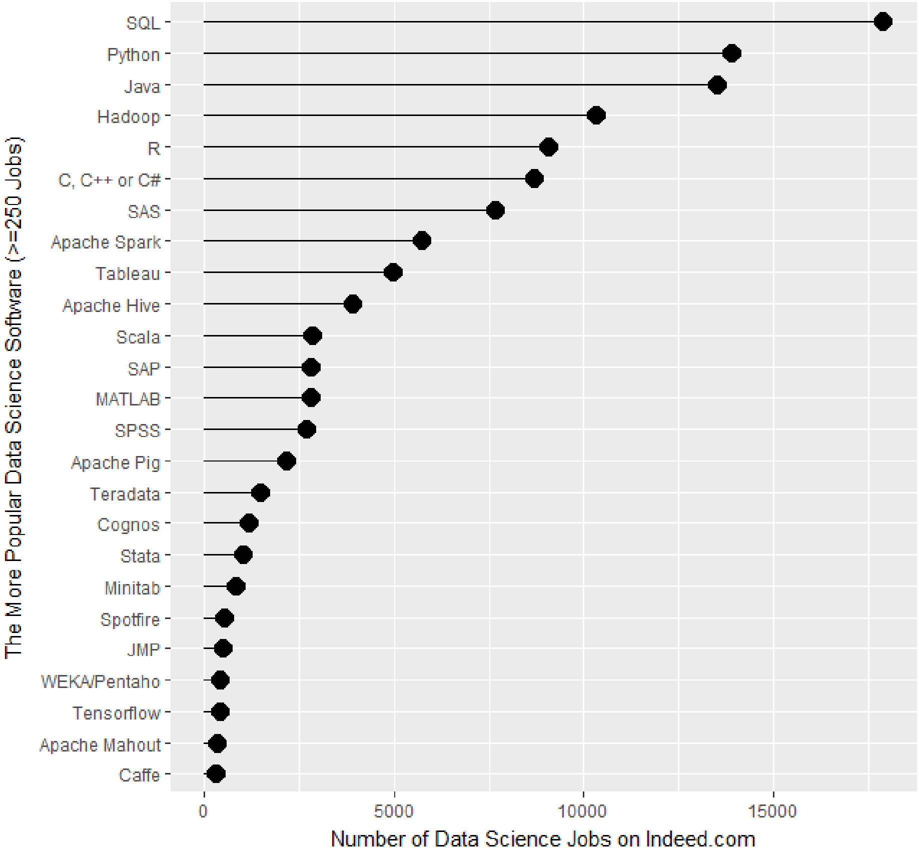


Figure 1. Programming Languages Sought on Indeed.com (2017)

Our graduate students get their introductory statistical training in “Statistics for Social Science” (POLS 581), a course I have taught since 2008. One of the learning objectives for this course is the ability to think critically in methodological terms, which includes writing programming scripts that run properly for themselves and others and contain comments explaining their work. (See Appendix A for the full list of learning objectives.) In the current rendition of this class, the students demonstrate their learning via lab homework, online quizzes, an in-class midterm, and two practical data projects resulting in programing scripts and papers. The transition to R requires changes to all of these components of the class in addition to lecture materials and handouts. These changes will be taking place regardless of awarding of the grant. If awarded, the grant will also allow us (i.e., myself and the teaching assistant) to create interactive tutorials in an R package called “Swirl” (see swirlstats.com) that would speed up the transition to R not only in POLS 581 but throughout the rest of the department, including current graduate students, faculty, and undergraduates.

In the current rendition of POLS 581, much time is spent in lab teaching students basic commands to generate statistical output and having them execute those commands. Less time is therefore spent working on substantive issues or working through practical data problems. Creating Swirl tutorials will allow us move the technical learning as work to be done before lab. (I also intend to have the students work through lessons in Lynda, such as “Data Wrangling in R”.)

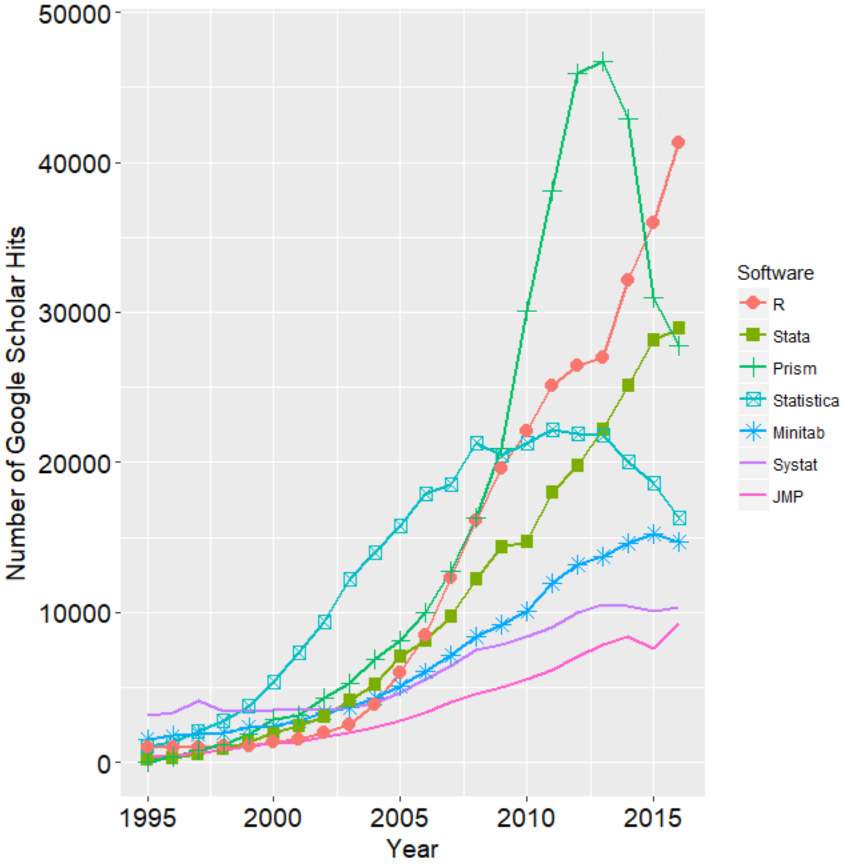


Figure 2. Statistical Software Usage in Articles (2017)

3 Student Backgrounds

4 Quizzing through the Ages

My first iteration of the class used more traditional teaching methods. In particular, I used pencil-and-paper quizzes taken during class. My grading philosophy focused on learning rather than evaluation. So, I would comment on the students' quizzes, return them, and, perhaps, discuss one question at the beginning of the next class that had given most students difficulty. My presumption was that the students would then continue their learning on their own.

Separately, I taught my first online class in the summer of 2012 where I learned the power and scope of the learning management systems and constructed module-level quizzes with deep question pools. As I mentioned above, I found that the quizzes could serve as learning tools. My consistent use of active learning attests to the fact that I believe we learn best by doing. So, by allowing multiple attempts with different flavors of questions that assessed the same learning objectives, the students could easily do electronically what I had always presumed they should be doing with the pencil-and-paper quizzes: taking them repeatedly until they really understood the material.

So, for the next two iterations of POLS 581, I used online quizzes with multiple attempts to encourage learning but gave an in-class exam that covered all the learning objectives from the previous exam and new material as well. I want to stress that the format of the exam remained pencil-and-paper; only the format of the quizzes were different. The assessment data here were clear and positive. For the learning objectives that were the same from the previous year, overall progress toward the same objectives scored 0.76 (out of 1.00, compared to the 0.65 the previous year); they now only scored less than 0.70 on four of the eighteen common objectives. The following year (2013), yet another cohort also scored 0.76 on these same objectives; they now scored less than 0.70 on only three of the eighteen common objectives. The quizzes were the major instructional change differentiating the 2011 cohort from the cohorts of 2012 and 2013.

While this was a considerable improvement, I'm always open to further improvement. It was about this time that I was introduced to gamified learning and the idea that I could encourage students to "level up" directly. Within the context of the online quizzes, this meant adding a bonus level for taking and scoring well on the quiz repeatedly. This was partly the genius of my graduate instructor at the time, Janelle Johnson, who had noted that when she scored a 100% on an initial quiz, she would call it a day. But then she was not as well prepared for that material for the in-class exam. Another aspect of creating an environment for safe interaction and repeated feedback is encouraging students who "got it" the first time around to retest their understanding. With this individual bonus, even students who scored well initially were encouraged to engage in repeated practice, which is generally known to aid learning. We also added a group bonus to encourage cohort building.

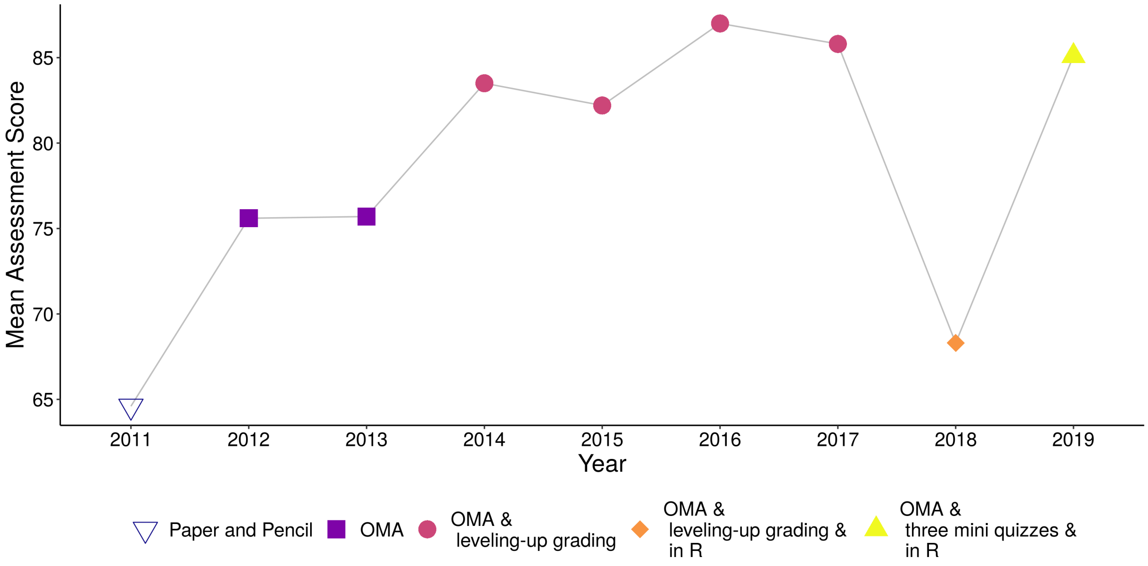


Figure 3. Benefits of Online Multiple Attempt (OMA) Quizzes

Once again, we gave a modified pencil-and-paper version of the midterm that included eighteen learning objectives that were measured the same going back to the 2011 exam. Also once again, we saw a jump in the assessment scores. Overall progress toward the common objectives scored 0.84 in 2014 and scored 0.82 in 2015. This had me hooked on this method.

There are four elements of this quizzing method that I felt needed tweaking. First, looking for at least three high scores among a student's multiple attempts for each quiz is both time intensive and not always easy for the student to keep track of. Second, a fair number of students didn't get the idea that repeated attempts were expected (either not realizing entirely or only considering it optional for bonus points). Third, it didn't explicitly encourage spaced practice as students could simply complete multiple attempts in quick succession. Fourth, it didn't allow for changes in difficulty or easily adding review questions as each weekly quiz drew from the same set of questions in the same proportions. (Finally but unrelated to the quizzing method itself, the quizzes had too many questions and presented a time burden.)

4.1 Interspacing and Interleaving

To deal with the problems with the quizzing method, we changed the weekly quizzes in Fall 2019. Now for each week's quiz, there were three stages (or mini-quizzes) containing five questions each. The first stage opened immediately after lecture ended, the second stage opened two days later, and the third stage opened four days later. The students could also progress to the next stage once they achieved a perfect score on the current stage. A student's grade for a weekly quiz would then be the simple sum of their best scores on each stage. Thus, they had a direct incentive to repeat a stage for which they hadn't yet gotten a perfect score and a direct incentive to take each stage at least once, enforcing practice.

While the weekly quiz tested the material covered in lecture, the first stage included questions from the previous week's quiz that the students demonstrated a need to re-examine. As the semester progressed, we would include review questions from more than just the previous week's quiz. We would also vary the questions from lecture across the stages with an intent to "ladder" more fundamental concepts in an early stage and derived concepts in a later stage (though we were not as systematic about this is we could have been).

Figures 4 through 6 present data on these quizzes from Fall 2019. Figure 4 shows that the Stage 1 mini-quiz for each weekly quiz generally had more attempts than the later stages. Week 4 and 7 are exceptions to this.

Figure 5 shows the distribution of attempt grades by stage. I expected higher and denser distribution for stages 2 and 3 compared to any given week's stage 1 (i.e., with a higher, tighter range than the previous stage and, thus, condensed in that range). However, there was much more variability, especially as the class started covering more mathematical concepts and the quizzes started including more calculation questions.

Figure 6 presents average grades over attempts for each stage of each weekly quiz. Trajectories such as those in panel e demonstrate what I expected within the data. Namely, students took stage 1 multiple times until they got a perfect score. Then they took the next stage. Having learned the material fairly well on stage 1, they needed fewer attempts to earn a perfect score on stage 2 (and similarly from stage 2 to stage 3). Other weeks show much more variability (perhaps especially Week 7 as shown in panel g where there were more attempts by most students for stages 2 and 3 than for stage 1).[4]

The trajectory in panel f presents an interesting pattern that students have told me about in previous years. A student getting a 4 out of 5 may recognize that they are having difficulty with one type of

[4] Indeed, this subfigure is actually truncated at 16 attempts. For Stage 2 of Week 7, the maximum number of attempts was 27.

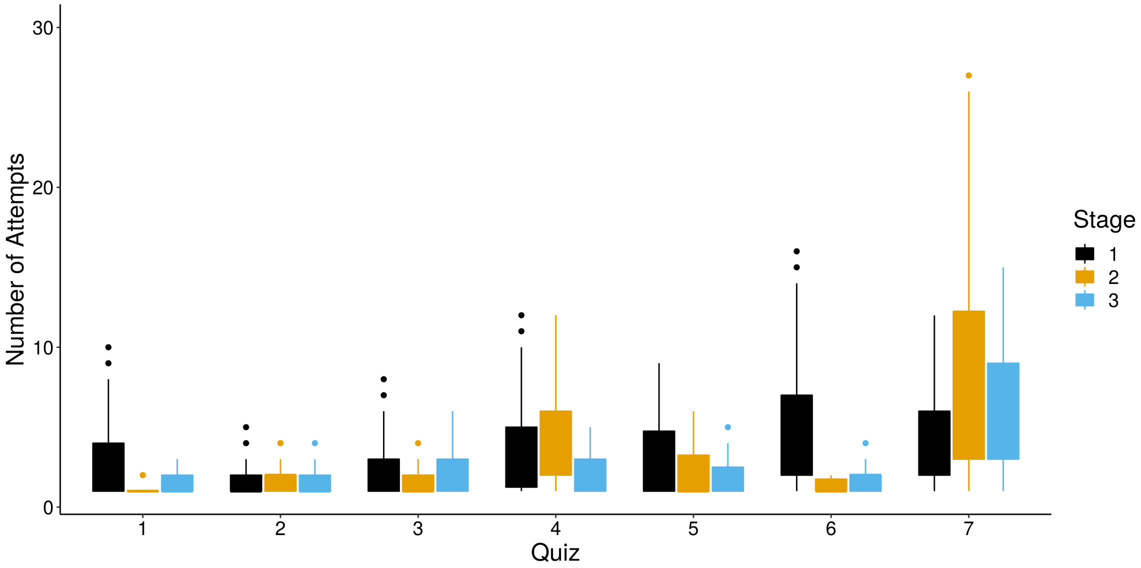


Figure 4. Box Plots of Number of Attempts on the Stages of Weekly Quizzes

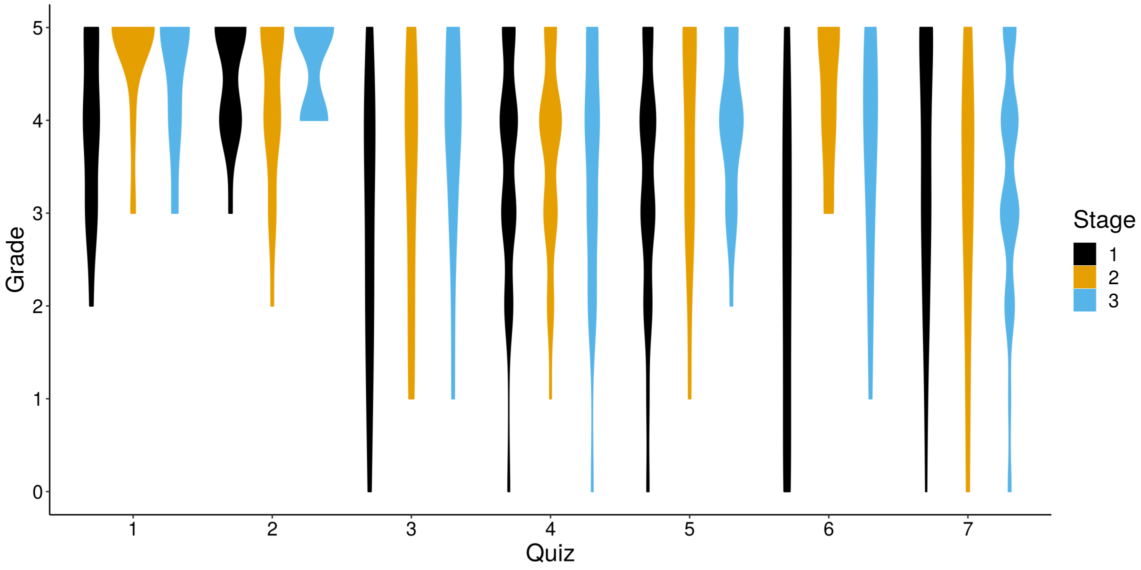


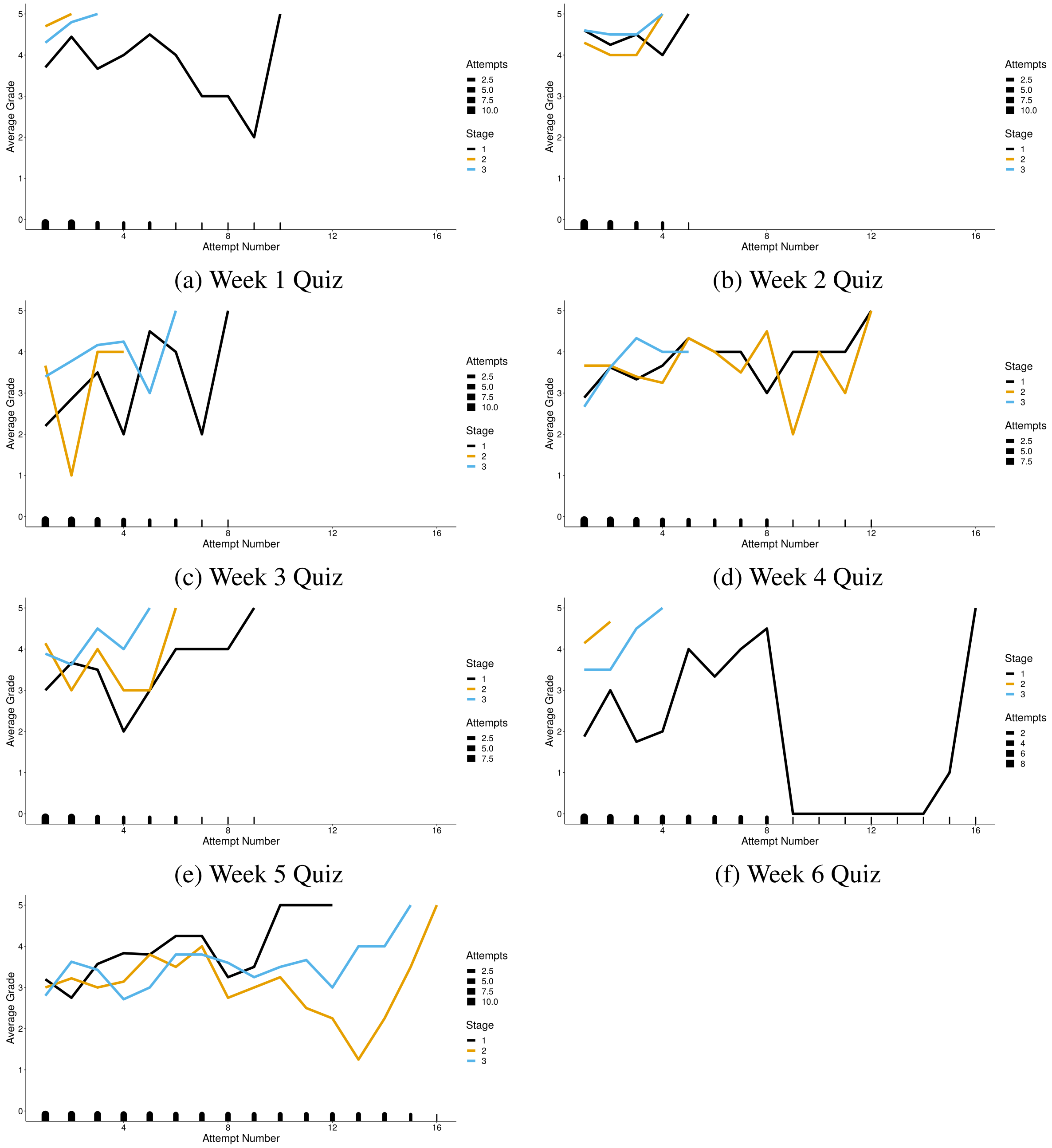
Figure 5. Distribution of Grades on the Stages of Weekly Quizzes

problem. Because their grade is the highest score for an attempt, they know there is no penalty for "taking" multiple attempts and only trying that one, difficult question—skipping the rest—until they finally (on attempt 15 in this case) get it right. On the next attempt, they answer all five questions hoping that this will be the last and best attempt (which it is on attempt 16). This exemplifies the student's own use of these quizzes for learning!

5 The Swirl Experiment

The title of my paper for the APSA 2020 TLC reflects an experiment to add even more automated teaching to my class. The punchline for this experiment is that it was too much, too fast. The following is from a teaching allocation grant proposal regarding our hopes for creating interactive tutorials using the Swirl library in R.

> The elements of a Swirl tutorial are text that would normally be part of lecture interspersed with automated checks for learning. These checks for learning include multiple-choice questions, value questions (entering the correct number), and expression questions (entering the correct formula or command). (Other types of questions are being developed by the Swirl community, including the ability to evaluate an entire programming script.) The tutorial doesn't allow the student to proceed until a question is answered correctly. For the expression questions, the students can pull up help on particular topics as well as create and manipulate data objects—all within the tutorial. This creates a double interactive feature in that a series of expression questions can step the student through an entire analysis of a particular data problem. For example, opening a data set, recoding variables, executing particular statistical tests, and having students identify values from the statistical output.
>
> The first lab session will focus on installing necessary software (including R, the user interface RStudio, the Swirl library, and the Swirl tutorials). The students will then go through the first couple Swirl tutorials (which will focus on understanding the user interface, beginning to understand R, and opening and summarizing data sets). After the first lab session, students will be expected to complete the Swirl tutorials before the lab session.



(a) Week 1 Quiz

(b) Week 2 Quiz

(c) Week 3 Quiz

(d) Week 4 Quiz

(e) Week 5 Quiz

(f) Week 6 Quiz

(g) Week 7 Quiz

Figure 6. Trajectory of Grades on the Stages of Weekly Quizzes

At the beginning of each subsequent lab session, we will have quick learning checks to make sure the students know the relevant commands that were covered in the tutorial. The remainder of each lab session will then present the students with a practical data problem that they should be able to complete during lab using the relevant commands. (For most of the current lab sessions, we have such homework problems. However, the students are rarely able to finish them during lab because so much time is spent teaching basic commands.) Flipping the lab sessions in this way will also open up more time for answering student questions as they work through the lab assignments.

The R programming language is uniquely equipped to facilitate innovative teaching approaches. All course modules will be available for student access from home and lab computers. Students will be able to practice programming syntax, be quizzed on statistical concepts, and review these materials repeatedly and with instant instructor-programmed evaluation and feedback (including pre-programmed hints). Because R and the Swirl lessons are platform independent, students will be free to engage in class review and homework from anywhere, instead of being tethered to the department computer lab. Finally, the portability of R and the Swirl lessons enable instructors to publish course materials online for review and use by other students and faculty across the world. Posting our tutorials on the Swirl repository will raise the visibility of the department and the University.

The portability of the Swirl tutorials implies that they can be used by an audience much larger

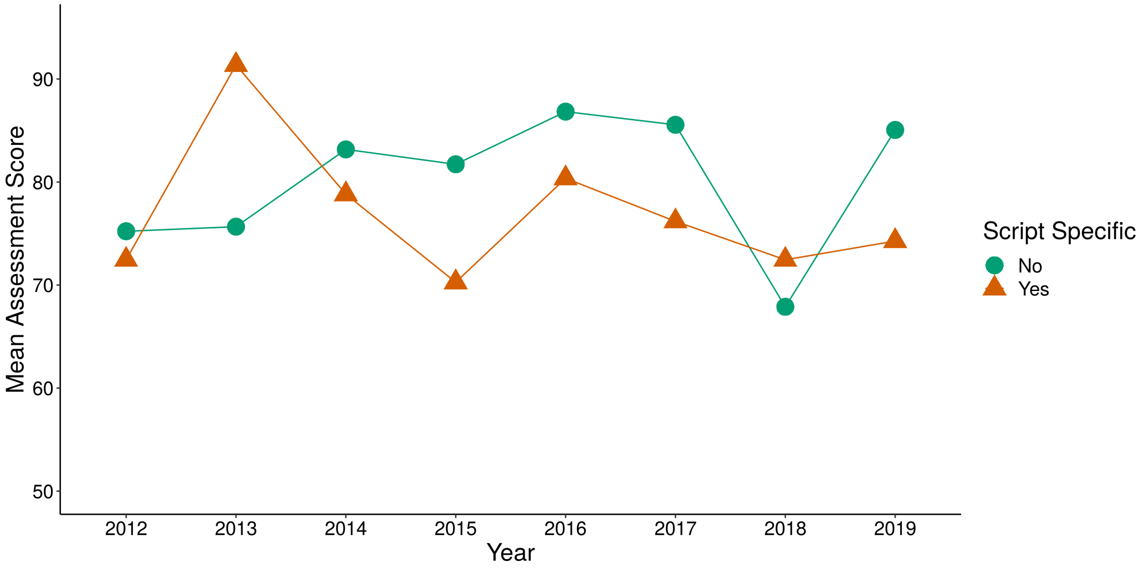


Figure 7. Midterm Assessment Disaggregated by Script Relevant or Not

> than the students in POLS 581. Our intention is to open these tutorials quickly to our current graduate students and faculty. This will provide a reliability check from those who already know R and kick start learning from those just getting started with R. As part of this, Colin Henry will lead a programming lab and working group to train graduate students who have already taken POLS 581 so they can more quickly transition from Stata to R. The programming lab and working group will rely on Swirl lessons adapted from the course materials as well as more extensive data management, function creation, and object-oriented programming training. It will provide a venue for collaborative research and opportunities to practice professional development and presentation skills.

While we did make progress on this front and we have tutorials accessible via GitHub (`https://github.com/ckbutler/POLS581_in_R`), we found more generally that we were hitting the students with too many different modes of presentation that weren't always on the same page. Sometimes this literally meant that we were using different libraries to have the students do the same statistical function. More fundamentally, Swirl presumes that you are teaching from the command line. This has its place, but—especially for graduate students—I focus on writing scripts that present complete demonstrations of a concept and how to modify it for their future purposes.

6 Starter Scripts

Transitioning to R has provided an opportunity to reflect on how I teach statistical programming. It has often been my goal to have the lab sessions be fairly self-guided. To that end, I took the time this last year to lead the lab sessions and write starter scripts for each (one in R and one in LaTeX) with task prompts embedded as comments. Once we got past the initial difficulties of installing software, learning about packages and libraries, and working with simple scripts, labs became routinized to the point that I was usually just helping students with individual difficulties.

7 Conclusion

I have two key takeaways from my experience teaching introductory methods. First, the online quizzes with unlimited attempts are an excellent learning tool for the students. Second, providing starter scripts to novice programmers worked well, but still requires working through concept laddering.

Appendices

A Learning Objectives

Students will demonstrate. . .

. . . (A) a firm understanding of research design and methods by

1. calculating various statistical values correctly;
2. correctly identifying the level of measurement of variables;
3. interpreting statistical tests appropriately;
4. summarizing data appropriately given the levels of measurement of the variables; and
5. executing appropriate statistical tests given the research question and levels of measurement of the variables.

. . . (B) an ability to think critically in methodological terms by

1. evaluating measurement validity;
2. evaluating testable hypotheses;
3. differentiating between theoretical concepts and measurable variables;
4. merging data sets appropriately given their units of analysis;
5. identifying and evaluating threats to statistical evidence;
6. evaluating the appropriateness of samples;
7. transforming variables appropriately given levels of measurement, the research hypothesis, and the appropriate test; and
8. writing scripts that run properly for themselves and others and contain comments explaining their work.

. . . (C) the capacity to conduct an original research design by

1. writing and justifying testable hypotheses;
2. drafting survey questions measuring political opinion; and
3. managing data collection for subsequent analysis.

. . . (D) analytical writing that is clear and appropriate to the audience by

1. connecting theoretical concepts to measurable variables;
2. writing and justifying testable hypotheses;
3. generating tables and figures for presentation;
4. interpreting statistical tests appropriately; and
5. analyzing theoretical arguments using statistical evidence.